# Observer Variability for Classification of Pulmonary Nodules on Low-Dose CT Images and Its Effect on Nodule Management[1]

Sarah J. van Riel, MD
Clara I. Sánchez, PhD
Alexander A. Bankier, MD, PhD
David P. Naidich, MD, PhD
Johnny Verschakelen, MD, PhD
Ernst T. Scholten, MD, PhD
Pim A. de Jong, MD, PhD
Colin Jacobs, MSc
Eva van Rikxoort, PhD
Liesbeth Peters-Bax, MD
Miranda Snoeren, MD
Mathias Prokop, MD, PhD
Bram van Ginneken, PhD
Cornelia Schaefer-Prokop, MD, PhD

**Purpose:** To examine the factors that affect inter- and intraobserver agreement for pulmonary nodule type classification on low-radiation-dose computed tomographic (CT) images, and their potential effect on patient management.

**Materials and Methods:** Nodules ($n$ = 160) were randomly selected from the Dutch-Belgian Lung Cancer Screening Trial cohort, with equal numbers of nodule types and similar sizes. Nodules were scored by eight radiologists by using morphologic categories proposed by the Fleischner Society guidelines for management of pulmonary nodules as solid, part solid with a solid component smaller than 5 mm, part solid with a solid component 5 mm or larger, or pure ground glass. Inter- and intraobserver agreement was analyzed by using Cohen κ statistics. Multivariate analysis of variance was performed to assess the effect of nodule characteristics and image quality on observer disagreement. Effect on nodule management was estimated by differentiating CT follow-up for ground-glass nodules, solid nodules 8 mm or smaller, and part-solid nodules smaller than 5 mm from immediate diagnostic work-up for solid nodules larger than 8 mm and part-solid nodules 5 mm or greater.

**Results:** Pair-wise inter- and intraobserver agreement was moderate (mean κ, 0.51 [95% confidence interval, 0.30, 0.68] and 0.57 [95% confidence interval, 0.47, 0.71]). Categorization as part-solid nodules and location in the upper lobe significantly reduced observer agreement ($P$ = .012 and $P$ < .001, respectively). By considering all possible reading pairs (28 possible combinations of observer pairs × 160 nodules = 4480 possible agreements or disagreements), a discordant nodule classification was found in 36.4% (1630 of 4480), related to presence or size of a solid component in 88.7% (1446 of 1630). Two-thirds of these discrepant readings (1061 of 1630) would have potentially resulted in different nodule management.

**Conclusion:** There is moderate inter- and intraobserver agreement for nodule classification by using current recommendations for low-radiation-dose CT examinations of the chest. Discrepancies in nodule categorization were mainly caused by disagreement on the size and presence of a solid component, which may lead to different management in the majority of cases with such discrepancies.



[1] From the Department of Radiology and Nuclear Medicine, Radboud University Nijmegen Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands (S.J.V.R., C.I.S., E.T.S., C.J., E.V.R., L.P.B., M.S., M.P., B.V.G., C.S.P.); Department of Cardiothoracic Imaging, Beth Israel Deaconess Medical Center, Boston, Mass (A.A.B.); Department of Radiology, NYU Langone Medical Center, New York, NY (D.P.N.); Department of Imaging and Pathology, Catholic University Leuven, Leuven, Belgium (J.V.); Department of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands (P.A.D.J.); and Department of Radiology, Meander Medical Center, Amersfoort, the Netherlands (C.S.P.). Received November 21, 2014; revision requested January 6, 2015; revision received February 6; accepted February 16; final version accepted March 17. Supported by a research grant from MeVis Medical Solutions AG, Bremen, Germany. **Address correspondence to** S.J.V.R. (e-mail: *sarah.vanriel@radboudumc.nl*).

Pulmonary nodules are the most frequent incidental findings in computed tomographic (CT) imaging of the chest. The advent of low-radiation-dose CT imaging for lung cancer screening further increased the detection of such nodules, and several societies published guidelines for their management strategies for intrapulmonary nodules (1–5). These guidelines use morphologic CT imaging criteria to estimate the risk of malignancy to trigger a management strategy, which includes no follow-up, follow-up with CT imaging after specific time intervals, and positron emission tomographic imaging, or invasive procedures, such as biopsy or resection. Although these guidelines differ with respect to their specific cut-off values for nodule size criteria, they uniformly differentiate between solid, part-solid, and pure ground-glass nodules. Because of the higher prevalence of invasive adenocarcinomas in part-solid nodules with a larger solid component (6–12), nodule management differs for part-solid nodules dependent on the size of the solid component (2–5).

Visual assessment of nodule morphologic structure on chest CT images, however, is prone to variability induced by the interpretation process. Manual diameter measurements as they are part of current management guidelines may further add to observer variability (13–15). Interobserver variability may affect patient management with potential effects on outcome, and will influence health care costs through the follow-up and work-up procedures.

## Advances in Knowledge

- Pulmonary nodule classification according to current guidelines has substantial inter- and intraobserver variability (mean κ, 0.51 and 0.57, respectively) on low-radiation-dose CT scans.
- Inter- and intraobserver variability is in almost 90% related to the definition and measurement of a solid component in pulmonary nodules.

The purpose of our study was therefore to examine the factors that affect inter- and intraobserver agreement for pulmonary nodule type classification on low-radiation-dose CT images. We analyzed the effects of features, which include anatomic location, size and nodule type, and the effect of image noise on observer agreement. Additionally, the potential effect on patient management was determined.

## Implication for Patient Care

- Inter- and intraobserver variability in nodule classification on low-radiation-dose CT scans results in potentially different nodule management strategies.

## Materials and Methods

### Materials

Nodules were selected from CT images from three sites of the Dutch-Belgian Lung Cancer Screening Trial (NELSON) (16). The trial was approved by the ethics committees of all participating centers and the Dutch Ministry of Health. Written informed consent was obtained from all participants at time of inclusion in the screening trial for acquiring the CT data and for analysis of these data for research purposes.

In NELSON, nodules detected by the screening radiologists were annotated by using size, nodule type (solid, part solid, or pure ground glass), and location as nodule features.

For our retrospective study, 160 unique pulmonary nodules were randomly selected (S.J.V.R. and C.S.P.) on the condition of an equal distribution of nodule types and similar size distribution. An uncontrolled random selection of nodules from the screening database would have led to an overload of small (<5 mm) solid nodules. To avoid this, the selection was performed randomly but with specific inclusion criteria, such as the nodule type, as noted in the screening database, and the nodule size.

The additional subcategorization of part-solid nodules with a solid component smaller than 5 mm and 5 mm or larger, respectively, was determined by the researcher (S.J.V.R.) on the basis of manual diameter measurements, averaged over length and width. The researcher underwent specific training and exclusively analyzed research on pulmonary nodules for 2 years. This resulted in 40 solid nodules, 40 pure ground-glass nodules, 40 part-solid nodules with a solid component 5 mm or larger, and 40 part-solid nodules with a solid component smaller than 5 mm. The 160 study nodules were located in 145 patients; 13 patients had two nodules and one patient had three nodules. The nodules were presented in random order and independently of each other to reduce interpretation bias. The nodule classification on the basis of the screening annotations did not serve as a reference standard, but was used during the inclusion process to ensure a relatively balanced distribution of the various nodule types. We used the diameters as reported in the NELSON database for the nodule selection process.

### CT Data Acquisition

The images used in our study were obtained between 2004 and 2010. All CT images in NELSON were acquired by

**Published online before print**
10.1148/radiol.2015142700 **Content codes:** CH CT

**Radiology 2015;** 277:863–871

**Abbreviation:**
NELSON = Dutch-Belgian Lung Cancer Screening Trial

**Author contributions:**
Guarantors of integrity of entire study, S.J.V.R., J.V., M.P., C.S.P.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.J.V.R., D.P.N., E.T.S., C.J., M.P., B.V.G., C.S.P.; clinical studies, S.J. V.R., J.V., P.D.J., L.P.B., M.S., B.V.G.; experimental studies, A.A.B., E.V.R., L.P.B., B. V.G.; statistical analysis, S.J.V.R., C.I.S., B.V.G., C.S.P.; and manuscript editing, S.J. V.R., C.I.S., A.A.B., D.P.N., J.V., P.D.J., C.J., E.V.R., M.S., M.P., B.V.G., C.S.P.

Conflicts of interest are listed at the end of this article.

using 16-detector–row CT scanners (Somatom Sensation 16, Siemens Medical Solutions, Forchheim, Germany; Mx8000 IDT or Brilliance-16P, Philips Medical Systems, Best, the Netherlands) and a low-radiation-dose protocol. Data were acquired by using 16 × 0.75 mm collimation, a tube current time product of 30 mAs, and a tube voltage of between 80 kVp and 140 kVp, dependent on the weight of the patient. Reconstructed section thickness was 1 mm, with a reconstruction increment of 0.7 mm. A moderately smooth reconstruction kernel was used (kernel B, Philips Medical Systems; and kernel B30f, Siemens Medical Solutions) (16,17).

### Observers

Eight radiologists from five institutions in Belgium, the Netherlands, and the United States participated as observers. Four were members of the Fleischner Society, two were involved in the NELSON trial readings, and two were general radiologists with a specific interest in thoracic radiology. Experience with interpretation of chest CT images ranged from 10 years to more than 30 years.

The radiologists were instructed to classify each nodule into one of the following four categories: solid, part solid with a solid component larger than or equal to 5 mm, part solid with a solid component smaller than 5 mm, or pure ground glass. Radiologists were free to use a caliper to determine the size of the solid component of a part-solid nodule. Radiologists were neither aware of the original classification in NELSON nor the number of nodules in each category.

Four radiologists interpreted the data set a second time after an interval of at least 12 weeks to assess intraobserver variability.

### Image Viewing

Nodules were presented in a random order. A dedicated reading display was used (Cirrus; Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands) to optimize reading efficiency. The CT data were loaded and displayed by using an enlarged view of each nodule in axial and coronal projection side by side. The radiologists could focus in or out, and change window settings as they deemed necessary. Interactive manual caliper measurements were available. Radiologists could scroll and review the complete CT examination if warranted.

### Nodule Characteristics and Image Quality

Four parameters were used to assess the effect of nodule characteristics and image quality on observer agreement: total nodule size (largest diameter in millimeters), nodule location (upper lobes, which included the lingula, lower lobes, or middle lobe), nodule type, and the presence of image noise in the lung parenchyma. The parameters nodule size, location, and type were extracted from the NELSON trial database. Image noise was measured by a researcher (S.J.V.R.) who did not participate in the reading process: Two 1-$cm^2$ regions of interest were placed in two homogeneous regions within the lung parenchyma close to the nodule, and the standard deviation of Hounsfield units averaged over the two measurements were the measure for image noise.

### Effect of Disagreement on Nodule Management

Effect on nodule management was on the basis of observer nodule classification, which was determined in the interpretation sessions and the original size measurements from the NELSON database. To evaluate the potential effect of observer variability on nodule management, we solely focused on two management strategies, namely CT examination follow-up (strategy I) or immediate work-up (strategy II), and did not further differentiate various follow-up delays. CT examination follow-up was assumed as the strategy for pure ground-glass nodules of any size, part-solid nodules with a solid component smaller than 5 mm, or solid nodules 8 mm or smaller. Immediate work-up by using additional tests was assumed as the strategy for part-solid nodules with a solid component 5 mm or larger and solid nodules larger than 8 mm.

Per nodule, we determined which strategy would have been potentially triggered by the radiologists' classifications. Because no reference standard was available, we determined per nodule the number of pair-wise strategy disagreements. Eight radiologists produced 28 reading pairs per nodule: Per reading pair, we determined whether a disagreement in classification, if present, would have potentially resulted in a different management strategy by taking two strategies into account, as earlier described. In addition, we determined per nodule the most frequent underlying morphologic structures that caused different classifications.

### Statistical Analysis

Multirater Fleiss κ statistics were used to measure interobserver agreement for nodule classification. Cohen κ statistics were applied to determine pair-wise inter- and intraobserver agreement. Pair-wise κ values were averaged over all possible observer pairs, which resulted in a mean with 95% confidence interval. κ values were interpreted by using the Landis and Koch guidelines (18). These statistics did not include original nodule annotations by the screening radiologists.

A multivariate analysis of variance was performed to assess the effect of nodule characteristics and image noise on observer agreement. Four nodule characteristics were considered: size, location, nodule type (solid, part solid, pure ground glass), and image noise. For nodule type classification we used the original screening annotations. Multivariate analysis of variance takes into account potential interactions between parameters (eg, lobar location and nodule type) and compensates for significance of interactions, if present.

Two subsets of nodules were defined according to the amount of agreement between radiologists: group A consisted of nodules with identical classification by at least seven of the eight radiologists and group B was composed of all remaining nodules. The subset (group A or group B) was the independent variable, and the four parameters (size, type, location, and noise) were the dependent variables.

**Table 1**

**Nodule Selection Characteristics**

| Nodule Annotation | No. of Nodules | Average Size (mm)* | Anatomic Location† | | | Sex† | |
|---|---|---|---|---|---|---|---|
| | | | UL | ML | LL | Male† | Female† |
| Solid | 40 | 9.2 (5.0–20.3) | 17 (42.5) | 8 (20) | 15 (37.5) | 37 (92.5) | 3 (7.5) |
| Part solid | 80 | 14.2 (5.0–33.0) | 54 (67.5) | 2 (2.5) | 24 (30) | 65 (81) | 15 (19) |
| With a solid component ≥5 mm | 40 | 17.1 (7.0–33.0) | 28 (70) | 1 (2.5) | 11 (27.5) | 31 (77.5) | 9 (22.5) |
| With a solid component <5 mm | 40 | 11.3 (5.0–21.0) | 26 (65) | 1 (2.5) | 13 (32.5) | 34 (85) | 6 (15) |
| Pure ground glass | 40 | 10.8 (5.0–26.4) | 24 (60) | 1 (2.5) | 15 (37.5) | 28 (70) | 12 (30) |
| All nodules | 160 | 12.1 (5.0–33.0) | 95 (59.3) | 11 (6.9) | 54 (33.8) | 130 (81) | 30 (19) |

Note.—LL = lower lobes, ML = middle lobes, UL = upper lobes including the lingula.

* Data in parentheses are range.

† Data in parentheses are percentages.

**Table 2**

**Inter- and Intraobserver Agreement in κ Values**

| Observer No. | Interobserver Agreement* | Intraobserver Agreement† |
|---|---|---|
| 1 | 0.57 (0.47, 0.66) | 0.71 (0.61, 0.78) |
| 2 | 0.56 (0.46, 0.65) | 0.56 (0.46, 0.67) |
| 3 | 0.53 (0.43, 0.63) | 0.55 (0.44, 0.64) |
| 4 | 0.53 (0.43, 0.62) | 0.47 (0.37, 0.56) |
| 5 | 0.55 (0.45, 0.64) | ND |
| 6 | 0.48 (0.39, 0.58) | ND |
| 7 | 0.45 (0.35, 0.54) | ND |
| 8 | 0.38 (0.29, 0.47) | ND |

Note.—Data in parentheses are 95% confidence intervals.

* κ averaged over all pair-wise κ values of each observer with the remaining seven observers.

† κ value of single observer.

We did not consider within-patient correlation among multiple nodules per patient because multiple nodules in a patient were considered to be individual nodules.

*P* values less than .05 were considered to indicate statistical significance. Analyses were performed by using statistical software (SPSS v. 20.0; SPSS, Chicago, Ill).

## Results

### Nodule Characteristics

The 160 nodules that were included in our study were an average size of 12.1 mm (range, 5–33 mm). Table 1 summarizes characteristics of the nodules as annotated in the screening database.

### Interobserver and Intraobserver Agreement

Interobserver agreement on nodule classification was moderate with a multirater Fleiss κ value of 0.50 (95% confidence interval: 0.48, 0.52).

The κ value between pairs of radiologists varied from 0.30 (fair agreement; 95% confidence interval: 0.22, 0.39) to 0.68 (substantial agreement; 95% confidence interval: 0.59, 0.76), with a mean of 0.51 (moderate agreement; 95% confidence interval: 0.41, 0.60) averaged over all radiologists.

Intraobserver agreement was moderate, with κ values ranging from 0.47 (moderate agreement; 95% confidence interval: 0.37, 0.56) to 0.71 (substantial agreement; 95% confidence interval: 0.61, 0.78), and a mean κ of 0.57 (95% confidence interval: 0.47, 0.66) averaged over four radiologists. Table 2 provides the results per radiologist.

### Nodule-Based Analysis

*Nodules with high observer agreement (group A).*—Group A consisted of 72 of the 160 nodules (45.0%), of which 45 (62.5%) were classified identically by all eight radiologists (Fig 1). The majority of these 45 nodules were solid (30 of 45 [66.6%]); 12 of 45 (26.6%) were part solid with a solid component 5 mm or larger, and the remaining three nodules were two pure ground-glass nodules and one part-solid nodule with a solid component smaller than 5 mm.

In 27 nodules (37.5%), one radiologist disagreed with the rest because of the presence of a solid component in 40.7% (11 of 27) or the size of this component in 44.4% (12 of 27). Most disagreement was caused by the same two radiologists (17 of 27 [63.0%]). The remaining 37.0% (10 of 27) of disagreements were caused by four more radiologists.

*Nodules with limited observer agreement (group B).*—Group B was composed of the remaining 88 nodules (55.0%), in which at least two radiologists disagreed with the rest. Disagreement involved only two categories in 35 of these 88 nodules (40.0%): In 11 cases, radiologists did not agree on the presence of a solid component within a nodule, and in 22 cases, radiologists disagreed whether the size of the solid component was smaller than 5 mm or 5 mm or larger.

In 42 nodules (47.7%), observer classifications varied between three categories. In all of these nodules, there was disagreement about the presence of a solid component, and in 27 cases there was additional disagreement about the size of this component relative to the 5-mm cutoff value.

In the remaining 11 nodules (12.5%), observer ratings varied between all four categories.

Details are provided in Table 3. Figure 2 shows examples of nodules with variable interobserver agreement.

### Morphologic Criteria and Image Quality

Multivariate analysis of variance found a significant difference between group A and group B (Pillai trace, 0.131; $F = 3.839$; $P = .001$) regarding the parameters nodule size, nodule location, nodule type (as classified by NELSON), and noise.

**Figure 1**

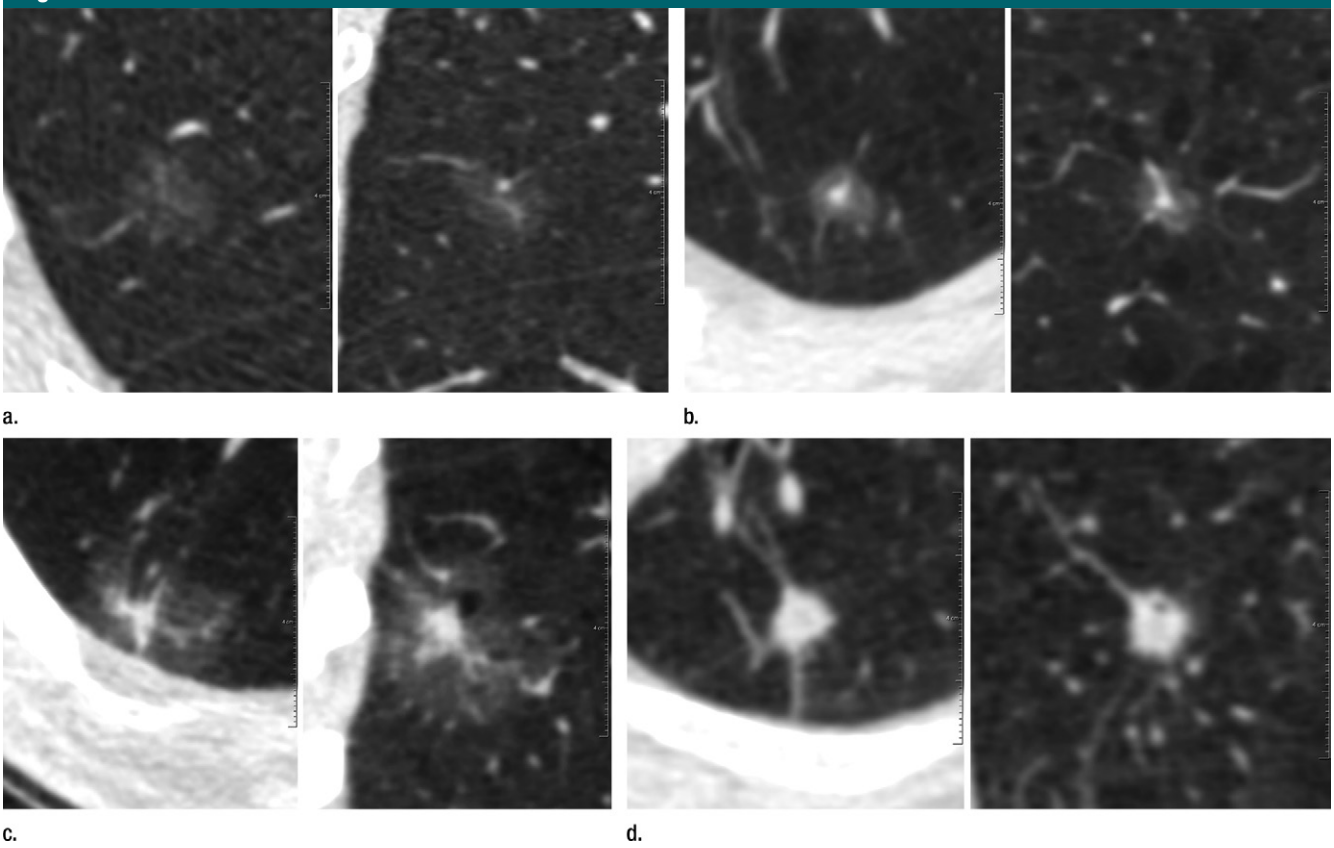


**Figure 1:** CT images show examples of nodules with complete agreement within all eight observers. Every nodule is displayed in axial view (left) and coronal view (right) by using lung windows. Each image shows a transverse field of view of 40 × 40 mm in which the nodule is centered. **(a)** Images show a nodule that is uniformly classified as pure ground-glass nodule. **(b)** Part-solid nodule is shown with a solid component smaller than 5 mm. **(c)** Part-solid nodule is shown with a solid component 5 mm or larger. **(d)** Images show a solid nodule.

The univariate $F$ tests revealed a significant difference for nodule location ($F$ = 6.463; $P$ = .012) and nodule type ($F$ = 15.862; $P < .001$). Nodules in group B were located in the upper lobes significantly more often and annotated as part-solid nodules compared to group A nodules. No significant difference was found for nodule size ($F$ = 2.426; $P$ = .121) or image noise by using continuous Hounsfield units ($F$ = 1.320; $P$ = .252). Table 4 summarizes these results.

### Effect of Observer Disagreement on Nodule Management

Grouping the radiologists' classifications according to management strategy I or II would have resulted in 96 nodules with at least one potentially different management strategy. In the remaining 19 nodules, different classification did not result in a potentially different management strategy.

In 17 of the 27 nodules from group A with disagreement of only one radiologist, the different classification would have resulted in a potentially different management strategy, and it was related to the presence and size of the solid component in 14 nodules.

In 26 of the 35 nodules from group B with two categories, classification would have resulted in potentially different management strategies, which was related to the solid component in 24 of the 26 nodules.

In all 42 nodules with three categories and 11 nodules with four categories, potential management differences would have occurred.

When all 28 possible reading pairs (28 possible combinations of observer pairs × 160 nodules = 4480 possible agreements or disagreements) were considered, disagreement about nodule type classification occurred in 36.4% (1630 of 4480), of which 88.7% (1446 of 1630) were related to the presence or size of a solid component. Two-thirds of the discrepant readings (1061 of 1630 [65.1%]) resulted in potentially different nodule management.

## Discussion

Several radiologic societies published recommendations for management of pulmonary nodules (1–5). They propose morphologic criteria on the basis of nodule type and size to determine follow-up intervals or further diagnostic

**Table 3**

**Disagreement Types in Subgroups A and B**

| Parameter | Group A | Group B |
|---|---|---|
| Complete agreement | | |
| SN | 30 (42) | … |
| PSN ≥5 | 12 (17) | … |
| PSN <5 | 1 (1) | … |
| GGN | 2 (3) | … |
| Disagreement about size of solid component between two categories | 12 (17) | 22 (25) |
| PSN ≥5 vs PSN <5 | | |
| Disagreement about presence of solid component between two categories, SN versus PSN or GGN versus PSN | 11 (15) | 11 (12.5) |
| Other disagreement between two categories, GGN versus SN | 4 (5) | 2 (2) |
| Disagreement about size and presence of solid component between three categories, PSN <5 versus PSN ≥5 and SN or GGN | … | 27 (31) |
| Disagreement about presence of solid component and other disagreement between three categories, SN or GGN versus PSN and SN versus GGN | … | 15 (17) |
| Disagreement between four categories, SN versus PS ≥5 versus PS <5 versus GGN | … | 11 (12.5) |

Note.—Data are number of nodules; data in parentheses are percentages. Group A consisted of 72 nodules with agreement of eight or seven observers. Group B consisted of 88 nodules with limited agreement between observers. SN = solid nodule, PSN = part-solid nodule, GGN = pure ground-glass nodule.

procedures. Part-solid nodules, in particular, are a focus of attention because of their high risk to represent malignancy (19). By following the guidelines, a more intense work-up of part-solid nodules is triggered beyond a certain (manually measured) diameter of the visually detected solid component.

In our study, we examined the factors that affected inter- and intraobserver agreement to retrospectively classify pulmonary nodules on low-dose CT images. We found a moderate overall interobserver agreement (mean κ, 0.51) to categorize nodules into solid, part solid with a solid component 5 mm or larger or less than 5 mm, and pure ground glass. While a high interobserver agreement was seen for solid nodules, the majority of disagreements related to either the presence of a solid component in part-solid nodules or the size of this solid component relative to the 5 mm threshold.

These results indicate that the evaluation of a potential solid component within a nodule that contained ground-glass components is prone to substantial interobserver variability. This variability is likely caused by the intrinsically subjective nature of the task in the absence of absolute measurement criteria. The Fleischner recommendations on management of subsolid nodules are adopted from the definitions in the Fleischner glossary: The solid component of a part-solid nodule had to fulfill the criteria of a consolidation, and the areas around it had to fulfill the definition of ground glass. These definitions are based on the degree of obscuration of the underlying lung architecture (2,20). On the basis of multiple studies (21–23), the Fleischner recommendations advise the use of mediastinal window settings to evaluate the solid component; lung window settings are recommended for assessment of the ground-glass component. However, it has to be noted that a widely used mediastinal window setting (width/length, 400/40) implies that only areas with densities that exceed −160 HU would be detected as a solid component. A recent study by Lee et al (6) based on correlation of CT morphologic and histopathologic analysis in a group of 59 part-solid nodules determined that a lower density range of −261 HU to −160 HU is most appropriate to describe the invasive tumor component.

The second major contributor to observer disagreement was the assessment of the size of the solid component. Several studies (6–12,24) found that the larger the solid component, the more likely the nodule will represent an invasive adenocarcinoma, which indicates a poorer prognosis. For this reason, current recommendations chose a size threshold that varies from 5 mm or larger to 8 mm or larger for the solid component to trigger a more aggressive work-up (2–5). Thus far, to evaluate the solid component, electronic caliper measurements based on the average of long and short axial dimensions in narrow and/or mediastinal windows were proposed (2). The results of our study confirm previous findings (13–15) that accuracy and reproducibility of manual caliper measurements of pulmonary nodules are limited. Clinical effect of such inaccuracy is the largest if measurements approach the decision threshold. It will largely affect reproducibility and thus standardization of classification, especially when several discriminating thresholds are on the basis of nodule size, which was proposed in the literature (4,5).

Our results are conformed to a recently published study (25) that also found a moderate interobserver variability for classification of subsolid nodules. Differences between this and our study are explicable by the size of the previous study group, the fact that the nodules were evaluated solely in axial scans at fixed window levels, and that the nodules were detected in clinical and not in screening studies. Nevertheless, the authors similarly noted that the presence of a solid component was a major contributor to variability.

Both the presence and the size of the solid component in a part-solid nodule represent decisive morphologic criteria to determine nodule management, which follows current recommendations. When a simplified dichotomous nodule management strategy is assumed, two-thirds of discordant readings resulted in different management decisions, and in the vast majority of these nodules, disagreement regarding the presence or size of the solid component was the reason for this conflicted strategy.

**Figure 2**

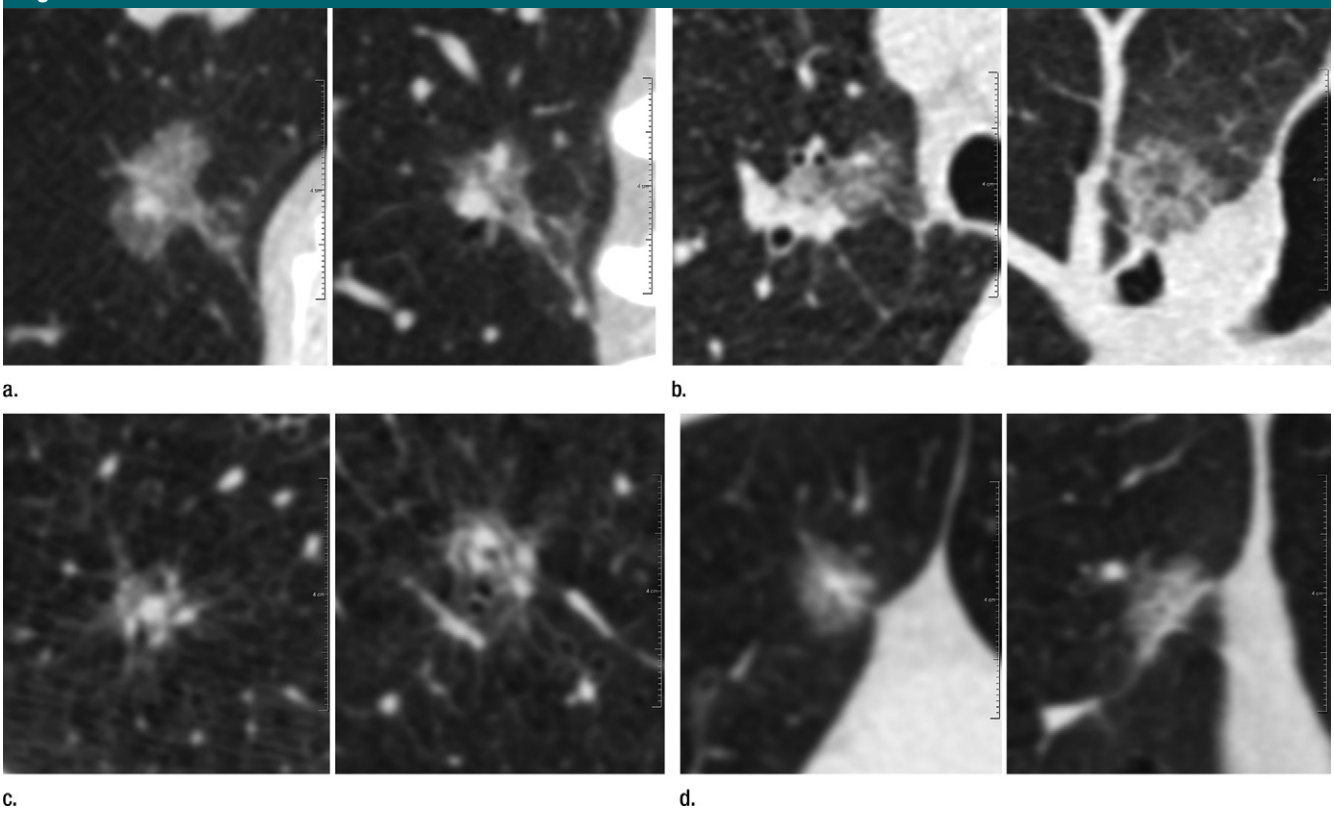


**Figure 2:** CT images show examples of nodules with varying nodule type classification disagreement scored by the observers. Every nodule is displayed in axial view (left) and coronal view (right) by using lung windows. Each image displays a transverse field of view of 40 × 40 mm, in which the nodule is centered. **(a)** In these CT images, four observers scored pure ground-glass nodule, three scored part-solid nodule with solid component less than 5 mm, and one observer scored part-solid nodule with solid component 5 mm or larger. **(b)** In these CT images, four observers scored pure ground-glass nodule, three scored part-solid nodule with solid component 5 mm or larger, and one observer scored part-solid nodule with solid component smaller than 5 mm. **(c)** In these CT images, four observers scored pure ground-glass nodule, two observers scored part-solid nodule with solid component 5 mm or greater, and two scored part-solid nodule with solid component less than 5 mm. **(d)** In these CT images, four observers scored part-solid nodule with solid component 5 mm or larger, three observers scored part-solid nodule with solid component smaller than 5 mm, and one observer scored pure ground-glass nodule.

In addition to nodule type, in our study, location in the upper lobe was found to have a statistically significant effect on observer disagreement. It must be noted that the statistics we used compensated for the fact that the prevalence of nodules was highest in the upper lobes. A potential explanation for this finding is that image artifacts are highest in the area of the shoulders. However, we did not find that image noise in the vicinity of the nodule had a statistically significant effect on variability.

The pair-wise agreement between radiologists showed a wide range of κ values. Some pairs of radiologists showed a substantial agreement while others showed only fair agreement. Intraobserver variability was moderate and similar to the average interobserver agreement, which suggested that disagreement was not only related to individual observer characteristics but also to nodule characteristics and the categorization task. We do not expect that training of radiologists would help because most of our readers were experienced with respect to analysis of nodules. Rather, we think that improvement is needed regarding the definition of the solid component in part-solid nodules. In the future, it is likely that automatic software will take over at least part of this task.

Our study has some limitations. Because observer variability is mainly caused by part-solid nodules, the relative number of these part-solid nodules will affect the overall observer variability in the cohort. We used an enriched cohort with similar size and numbers of nodules within the various nodule categories. We chose this approach to compensate for the disproportionately large number of small solid nodules in a screening population to make the best use of observer time for our particular study. The simplified assumption of two management strategies did not consider any other morphologic characteristics (eg, spiculation) for further risk estimation, as suggested in the literature (4). However, we believe that

**Table 4**

**Effect of Nodule Location, Type, Image Noise, and Nodule Size on Observer Agreement**

| Parameter | Group A | Group B | *P* Value |
|---|---|---|---|
| Nodule location | | | |
| Upper lobe | 35 (49) | 60 (68.2) | .012 |
| Middle lobe | 8 (11) | 3 (3.4) | … |
| Lower lobe | 29 (40) | 25 (28.4) | … |
| Nodule type | | | <.001 |
| Solid | 36 (50) | 4 (4.5) | … |
| Part solid | 17 (24) | 63 (71.5) | … |
| Pure ground glass | 19 (26) | 21 (24) | … |
| Average noise (HU) | 38.3 | 41.7 | .252 |
| Average nodule size (mm) | 11.3 | 12.7 | .121 |

Note.—Data are number of nodules unless otherwise indicated; data in parentheses are range. *P* values were derived from univariate *F* test. Group A consisted of 72 nodules with agreement of eight or seven observers. Group B consisted of 88 nodules with limited agreement between observers.

the interobserver variability in our study for the group of part-solid nodules was realistic and will translate into potentially differing management strategies, as described.

Furthermore, all CT images were acquired with low-radiation-dose technique by using moderately smoothing reconstruction kernels to improve the signal-to-noise ratio. High-resolution kernels are frequently used for the lung. They provide higher spatial resolution and are therefore recommended in current recommendations for nodule management (2,3). However, they also substantially increase image noise with negative effects on detail visibility and therefore were not used for reconstruction of the low-dose CT data in our study. It remains uncertain whether at all and to what extent nodule evaluation would have benefited from a high-spatial-resolution kernel in these low-dose images.

Finally, a histopathologic reference standard was not available for this data set. Our study does not focus on the prediction of correct management; it focuses on the effect of observer variability on management decisions. For this goal, no reference standard is required.

In summary, we found moderate observer agreement for nodule classification by using current recommendations. Discrepancies were mainly caused by disagreement in the size and presence of a solid component in part-solid nodules, which led to potentially different management in the majority of cases.

**Acknowledgments:** We thank the investigators of the NELSON trial for providing data for this study.

**Disclosures of Conflicts of Interest: S.J.V.R.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author reported a research grant from Mevis Medical Solutions AG. Other relationships: disclosed no relevant relationships. **C.I.S.** disclosed no relevant relationships. **A.A.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author received payment for a consultancy from Spiration, Olympus; author received payment for lectures from American Thoracic Society; author received payment for royalties from Elsevier. Other relationships: disclosed no relevant relationships. **D.P.N.** disclosed no relevant relationships. **J.V.** disclosed no relevant relationships. **E.T.S.** disclosed no relevant relationships. **P.A.D.J.** disclosed no relevant relationships. **C.J.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author reported a research grant from Mevis Medical Solutions AG. Other relationships: disclosed no relevant relationships. **E.V.R.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author receives payment for employment and stock options from Thirona BV. Other relationships: disclosed no relevant relationships. **L.P.B.** disclosed no relevant relationships. **M.S.** disclosed no relevant relationships. **M.P.** Activities related to the present article: author's institution receives payments for a grant from Toshiba Medical Systems. Activities not related to the present article: author receives payment for lectures from Bayer, Toshiba Medical Systems, and Bracco. Other relationships: disclosed no relevant relationships. **B.V.G.** disclosed no relevant relationships. **C.S.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author received travel expenses from Riverain Technologies. Other relationships: disclosed no relevant relationships.

## References

1. MacMahon H, Austin JH, Gamsu G, et al. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 2005;237(2):395–400.
2. Naidich DP, Bankier AA, MacMahon H, et al. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. Radiology 2013;266(1):304–317.
3. Manos D, Seely JM, Taylor J, Borgaonkar J, Roberts HC, Mayo JR. The Lung Reporting and Data System (LU-RADS): a proposal for computed tomography screening. Can Assoc Radiol J 2014;65(2):121–134.
4. Lung-RADS Assessment Categories. Version 1.0. American College of Radiology. Lung CT Screening Reporting and Data System (Lung-RADS) Web site. http://www.acr.org/Quality-Safety/Resources/LungRADS. Published April 28, 2014. Accessed September 15, 2014.
5. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013;143(5 Suppl):e93S–e120S.
6. Lee KH, Goo JM, Park SJ, et al. Correlation between the size of the solid component on thin-section CT and the invasive component on pathology in small lung adenocarcinomas manifesting as ground-glass nodules. J Thorac Oncol 2014;9(1):74–82.
7. Ohde Y, Nagai K, Yoshida J, et al. The proportion of consolidation to ground-glass opacity on high resolution CT is a good predictor for distinguishing the population of non-invasive peripheral adenocarcinoma. Lung Cancer 2003;42(3):303–310.
8. Nakata M, Sawada S, Yamashita M, et al. Objective radiologic analysis of ground-glass opacity aimed at curative limited resection for small peripheral non-small cell lung cancer. J Thorac Cardiovasc Surg 2005;129(6):1226–1231.
9. Aoki T, Tomoda Y, Watanabe H, et al. Peripheral lung adenocarcinoma: correlation of thin-section CT findings with histologic prognostic factors and survival. Radiology 2001;220(3):803–809.
10. Kim EA, Johkoh T, Lee KS, et al. Quantification of ground-glass opacity on high-resolution CT of small peripheral adenocarcinoma of the lung: pathologic and prognostic implications. AJR Am J Roentgenol 2001;177(6):1417–1422.
11. Kobayashi N, Toyooka S, Ichimura K, et al. Non-BAC component but not epidermal growth factor receptor gene mutation

is associated with poor outcomes in small adenocarcinoma of the lung. J Thorac Oncol 2008;3(7):704–710.

12. Vazquez M, Carter D, Brambilla E, et al. Solitary and multiple resected adenocarcinomas after CT screening for lung cancer: histopathologic features and their prognostic implications. Lung Cancer 2009; 64(2):148–154.
13. Gierada DS, Pilgram TK, Ford M, et al. Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. Radiology 2008;246(1): 265–272.
14. Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, Engelke C. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. Eur Radiol 2006;16 (4):781–790.
15. Singh S, Pinsky P, Fineberg NS, et al. Evaluation of reader variability in the interpretation of follow-up CT scans at lung cancer screening. Radiology 2011;259(1):263–270.
16. van Klaveren RJ, Oudkerk M, Prokop M, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361(23):2221–2229.
17. Ru Zhao Y, Xie X, de Koning HJ, Mali WP, Vliegenthart R, Oudkerk M. NELSON lung cancer screening study. Cancer Imaging 2011;11(Spec No A):S79–S84.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
19. Henschke CI, Yankelevitz DF, Mirtcheva R, et al. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. AJR Am J Roentgenol 2002;178(5): 1053–1057.
20. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology 2008;246(3):697–722.
21. Kakinuma R, Kodama K, Yamada K, et al. Performance evaluation of 4 measuring methods of ground-glass opacities for predicting the 5-year relapse-free survival of patients with peripheral nonsmall cell lung cancer: a multicenter study. J Comput Assist Tomogr 2008;32(5):792–798.
22. Takamochi K, Nagai K, Yoshida J, et al. Pathologic N0 status in pulmonary adenocarcinoma is predictable by combining serum carcinoembryonic antigen level and computed tomographic findings. J Thorac Cardiovasc Surg 2001;122(2):325–330.
23. Okada M, Nishio W, Sakamoto T, Uchino K, Tsubota N. Discrepancy of computed tomographic image between lung and mediastinal windows as a prognostic implication in small lung adenocarcinoma. Ann Thorac Surg 2003;76(6):1828–1832; discussion 1832.
24. Park CM, Goo JM, Kim TJ, et al. Pulmonary nodular ground-glass opacities in patients with extrapulmonary cancers: what is their clinical significance and how can we determine whether they are malignant or benign lesions? Chest 2008;133(6):1402–1409.
25. Penn A, Ma M, Chou BB, Tseng JR, Phan P. Inter-reader variability when applying the 2013 Fleischner guidelines for potential solitary subsolid lung nodules. Acta Radiol 2014 Oct 7. [Epub ahead of print]